Revolutionizing Company Perceptions: Dual-Agent Architecture Powering AI-Driven Likert Scale Systems for Organizational Ratability

Kemal Abdullah, Dr.Noorhan Abbas

University of Leeds, Leeds, UK.

Abstract. Accurately capturing and quantifying employee sentiment remains a central challenge in organizational research. Traditional methods often lack the nuance to reflect complex perceptions, prompting the rise of machine learning (ML) as a transformative tool in sentiment analysis. This study presents a dual-agent AI framework that leverages two large language models (LLMs): DeepSeek-R1, which generates authentic employee reviews, and GPT-4o serves as the critic sharpens the output through an iterative, dynamic discourse. To enhance efficiency, we integrate a teacher-student paradigm, distilling DeepSeek-R1 generated outputs into more efficient models—LLaMA and Qwen—paired with DeBERTa-v3 large and RoBERTa larg for classification. Our experiments identify LLaMA and DeBERTa-v3 large as the optimal combination, achieving 55.43% accuracy and 55.73% F1 score, significantly surpassing our base model BiLSTM. SHAP values are employed to validate classification decisions, promoting transparency and fairness in organizational assessments. With sentiment mapped to a 5-point Likert scale, this framework not only advances data-driven decision-making but also offers broad applicability across education technology, marketing, and beyond—paving the way for interdisciplinary innovation.

Keywords: Feedback Sentiment Analysis, LLMs, Model Distillation, Transfer Learning, Explainable AI

1 Introduction

Employee perspectives are central to modern organizational research[11] Systematic metrics provide reliable frameworks to assess perceptions and performance in dynamic environments, with sentiment analysis emerging as a key tool for extracting insights from textual data. The 5-point Likert scale[6] balances simplicity and depth, enabling nuanced sentiment quantification in organizational studies. However, qualitative feedback's variability and noise challenge structured analysis, demanding innovative solutions[16]. Evolving from rudimentary methods, modern sentiment analysis integrates advanced technologies[14]. This study leverages LLMs on diverse English-language datasets using the 5-point Likert scale to standardize company perception evaluations, bridging methodological gaps. The approach ensures adaptability across linguistic and cultural contexts while enhancing sentiment granularity.

Inspired by the innate and natural process of human reasoning and drawing from the iterative *Self-Refinement* principles seen in[12]. Our framework orchestrates a dialogue between two distinct yet harmonized Agents. DeepSeek-R1 (Composer) crafts insightful reviews, while GPT-4o (Critic) refines them iteratively[10]. Expanding the collaborative dual-agent framework into a dynamic learning ecosystem, we introduce a teacher-student paradigm, DeepSeek-R1 (Teacher), processes data, offering in-depth reviews that are distilled and condensed into a smaller, more efficient LLaMA and Qwen models (Student)[4, 7]. These are paired with DeBERTa-v3 large and RoBERTa large classifiers, creating a robust classification framework. This iterative approach enhances reasoning capabilities and advances sentiment analysis, generating nuanced yet actionable insights. These outcomes enable better informed decisions while accelerating interdisciplinary research innovation.

2 Literature Review

Employee sentiment profoundly impacts a company's perception and image. Cutting-edge algorithms have reshaped employee sentiment analysis through binary classification systems. For instance, the hybrid approach, such as an Auto-encoder paired with Support Vector Machine (AE-SVM)[22] achieved an F1 score of 86.3% in analyzing regional employee reviews (e.g., Guangdong's leather industry), though its narrow geographic scope hinders broader applicability. Conversely, Term Frequency-Multilayer Perceptron (TFIDF-MLP) [9] attained 96.74% accuracy on Kaggle's Glassdoor dataset. Despite this success, the study highlighted the needs for advanced encoding and hyper-parameter optimization to maximize performance.

Staying on the binary classification, Yadav et al.[21] analyzed Google employee Glassdoor reviews using ML techniques, with Random Forest (RF) achieving 88.07% accuracy, future work proposes refining linguistic subtleties and ordinal sentiment classes. Ke et al.[8] developed Employee Sentiment Analysis and Management System (ESAMS), a real-time monitoring tool using a Naive Bayes (NB) classifier that achieved 74% accuracy in a small-scale pilot, though it was limited by brief testing, limited candor, and reliance on SnowNLP. Similarly, Bajpar et al.[2] employed an Extreme Learning Machine (ELM)-Doc2vec ensemble to generate 300-dimensional embeddings from Glassdoor reviews (74.89% accuracy), with future aims to integrate employee ratings into a product-sentiment model and enhance aspect-level analysis for deeper insights.

Prior studies prioritized binary sentiment classification efficient but lacking nuance, often missing sarcasm and emotional depth, underscoring the need to adopt advanced frameworks. Fouda et al.[5] addresses this by introducing neutrality to Arabic specific analysis (108k reviews), employing tailored ML, Logistic Regression (LR) achieves 94% accuracy, though linguistic nuance loss persists, suggesting future Deep Learning (DL) integration. Rehan et al.[15] fused numerical ratings and textual analysis using an Extra Tree Classifier (ETC)-100% for ratings and 79% for text—identifying 76% of reviews as "proper" with future aims to apply DL and topic modeling uncover dissatisfaction drivers. Meanwhile, Abid et al.[1] expanded sentiment tiers to five categories (Super Positive/Super Negative), employing a Bidirectional Long Short-Term Memory (BiLSTM)–Artificial Neural Network (ANN) model on Glassdoor reviews from top Bangladeshi IT firms and Amazon product reviews for enhanced dataset diversity, achieving an F1 score of 88.64%. However, the study's narrow geographic and industry focus limits broader applicability, emphasizing the necessity for diverse datasets

Beyond employee sentiment analysis, cross-domain methodologies offer critical insights. Two studies[18, 3] employ ML to predict product ratings from customer reviews. The first uses multi-class with LR model, achieving 54.1% accuracy, with plan to expand by incorporating review dates, analyzing sales impact, and adapting the framework for other platforms. The second achieves 96% accuracy via a Convolutional neural network (CNN)-Recurrent Neural Network (RNN)-BiLSTM ensemble and 87.5% with RNN alone. Limitations in language and fake review handling reduce the model's effectiveness in multilingual, diverse settings. Focusing further on product review analysis, Wang et al.[19] propose Dynamic TextRank-Feature Fusion for Long Short-Term Memory-Feed Forward Neural Network (DTFS-LFNN), a lightweight psychology-driven model for small e-commerce platforms. Combining refined TextRank with LSTM-Feed Forward Neural Network (FFNN) ensemble learning, it achieves 89.5% accuracy and F1/F2 scores of 0.799/0.765, with future work targeting enhanced ensemble methods and parameter optimization. Similarly, Roumetiotis et al.[17] benchmark GPT-3.5, LLaMA-2, Bidirectional Encoder Representations from Transformers (BERT), and RoBERTa for e-commerce sentiment analysis, with GPT-3.5 achieving the highest accuracy 64.24%. However, the 5,029 review dataset lacked diversity representation, urging broader data integration and ethical considerations like privacy.

3 Methodology

3.1 The Collaborative Dual Agent Framework: An Architectural Overview

Recognizing the risk of inductive bias, we utilize a dual-agent framework featuring distinct models to counteract and mitigate its effects[10]. DeepSeek-R1, selected as the Composer for surpassing all open-source models and rivaling leading closed-source counterparts, crafts synthetic company reviews by weighing two variables, the pros and cons of employee feedback through a Chain-of-Thought (CoT) paradigm. This approach enhances dataset quality by embedding nuanced perspectives and minimizing noise. The Critic, powered by GPT-4o, a versatile general purpose model chosen for its adaptability and evaluative precision—systematically assesses coherence, holistic employee feedback, and readability through iterative scoring until strict quality benchmarks are achieved. Through comparative experimentation, DeepSeek-R1 outperformed DeepSeek-V3 as the Composer, while GPT-4o with a temperature of 0.5 demonstrated optimal performance as the Critic. Refined reviews are systematically paired with original feedback, ensuring dataset consistency. See Fig. 1 depicts our framework.

3.2 Fine-Tuning LLMs for Efficient Model Distillation and Classification

Faced with the challenge of real-time inference. we take advantage of model distillation, transforming the iterative dual-agent framework into a single, computationally efficient model for one-shot review generation. Central to this study is Qwen, a versatile model trained on extensive web data and a multilingual corpus across 29+ languages. Its strong performance in coding, mathematics, and text generation—combined with its cross-lingual capabilities—makes it especially effective for English sentiment analysis. In contrast, LLaMA was trained on diverse public sources, such as Wikipedia and

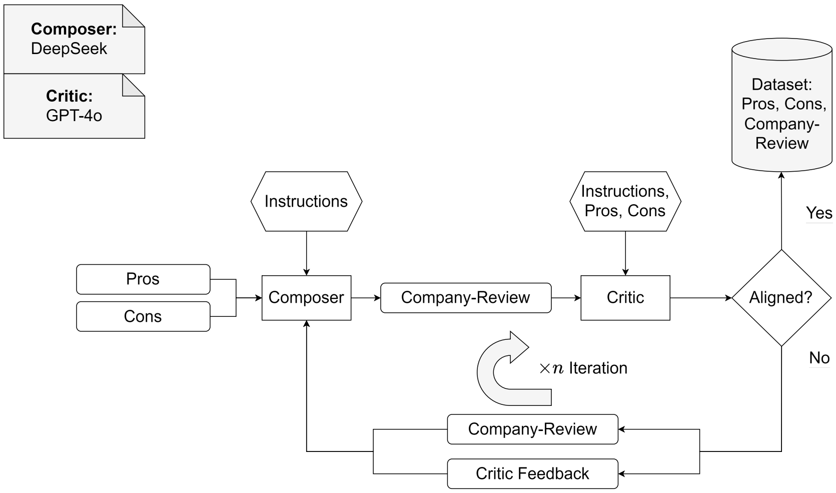


Fig. 1: The architecture described above outlines the collaborative framework between two agents - Composer and Critic- to improve a generated company review

CommonCrawl, spanning eight languages, and was specifically designed to excel in assistant-like chat tasks, While this aligns well with our objectives, its limited multilingual range compared to Qwen sets the stage for a comparative analysis, highlighting their unique strengths and trade-offs.

The distillation process involved fine-tuning models of varying sizes-Qwen 2.5 (7B, 1.5B) , LLaMA 3.1 8B, and LLaMA 3.2 1B-using Supervised Fine-Tuning (SFT) with LoRa (Low-Rank Adaptation) for efficiency. Training models of varying sizes serves multiple purposes: larger models harness their extensive parameters to capture complex linguistic patterns and contextual dependencies, while, smaller ones offer computational efficiency and adaptability for resource-constrained environment, allowing a balanced evaluation of performance versus efficiency. Additionally, quantized versions of LLaMA 3.1 8B and Qwen 2.5 7B were also explored to assess computational trade-offs. All models were benchmarked using BERTScore. To enhance interpretability to distilled reviews, we introduced a rating classification mechanism that maps distilled outputs into a 1–5 scale. Leveraging DeBERTa-v3 large, renowned for its cutting-edge architecture—featuring Disentangled Attention and Enhanced Mask Decoder—we prioritized its use as the most advanced transformer-based classifier[1], while RoBERTa-large served as a benchmark for a comparative analysis. Both models were trained on distilled outputs from LLaMA 3.1 8B and the quantized version of LLaMA 3.1 8B, with their performances was assessed using Accuracy and F1-score metrics. View the full pipeline in Fig. 2.

3.3 Optimizing Hyper-parameters for Effective Fine-tuning

The conventional hyper-parameters from Hugging Face did not yield optimal results, necessitating hyper-parameter tuning, which served to optimizing performance across

[1] https://github.com/microsoft/DeBERTa

tasks. In model distillation, LoRA rank and LoRA alpha balanced efficiency and adaptability, while epochs and batch size influenced training depth and memory usage. The learning rate, adjusted through warmup steps and a scheduler, ensured stable weight updates, with gradient check-pointing improving memory efficiency. Similarly, for classification, epochs and batch size were critical for effective training, while the learning rate, guided by a scheduler and warmup ratio, stabilized early training and fine-tuned weight adjustments. Collectively, these hyper-parameters structured a robust optimization process that maximized performance across both tasks

3.4 Performance Metrics

For model distillation effectivenes, we employed BERTScore, which measures semantic similarity between generated and reference texts via token embeddings (x, y):

$$sim(x, y) = x^{\intercal} y \tag{1}$$

BERTScore calculates token-level cosine similarities, aggregating them via precision, recall, and F1-score at the embedding level. Precision (P) and Recall (R) are defined as:

$$P_{\text{BERT}} = \frac{1}{|Y|} \sum_{y \in Y} \max_{x \in X} sim(x, y) \quad (2) \qquad R_{\text{BERT}} = \frac{1}{|X|} \sum_{x \in X} \max_{y \in Y} sim(x, y) \quad (3)$$

Finally, the BERTScore F1 metric is computed as:

$$F_{\text{BERT}} = 2 \cdot \frac{P \cdot R}{P + R} \tag{4}$$

For classification. Accuracy reliably measures correct prediction in our balanced dataset:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(y_i = \hat{y}_i) \quad \text{where} \quad \mathbb{I}(y_i = \hat{y}_i) = \begin{cases} 1 & \text{if } y_i = \hat{y}_i, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The macro-average F1 offers additional insights by averaging per-class performance:

$$\text{Macro-average F1} = \frac{1}{n} \sum_{i=1}^{n} F1_i \tag{6}$$

$$\text{where} \quad F1_i = 2 \cdot \frac{\text{precision}_i \cdot \text{recall}_i}{\text{precision}_i + \text{recall}_i} \quad \text{for class } i. \tag{7}$$

Here, Precision measures the ratio of true positives to total predicted positives,

$$\text{precision}_i = \frac{TP_i}{TP_i + FP_i}, \tag{8}$$

While, Recall measures the ratio of true positives to all actual positives.

$$\text{recall}_i = \frac{TP_i}{TP_i + FN_i}. \tag{9}$$

Combining BERTScore with Accuracy and F1-score enables a well-rounded evaluation of semantic quality and predictive performance.

3.5 Dataset

The foundation of our study was a publicly available Glassdoor dataset from Kaggle[2], which we later found to be plagued by class imbalances and a narrow diversity spectrum. To ensure a more equitable dataset, we extended our data collection via web scraping from Glassdoor[3] and AmbitionBox[4], a leading Indian review platform, leveraging *hrequests* for seamless data acquisition and *BeautifulSoup4* for precise parsing. This data augmentation aims to reduces bias while expanding the model's capacity to capture diverse organizational and employee experiences. Targeted refinements were applied to optimize dataset quality and framework compatibility:

Feature Selection To ensure robust feature selection, Pearson and Kendall correlation tests were used to assess relationships between variables, revealing weak associations (range: 0.035–0.50), leading to the removal of nine features, including work-balance-stars and career-opportunities-stars. Company and Location were also removed to improve model generalizability, aligning with the study's focus on employee feedback.

Data Preprocessing Employee feedback was systematically refined, ensuring linguistic consistency and contextual integrity. The core preprocessing steps are:

1. Handling Missing Values: We replace null entries in features like *Pros* and *Cons* with a standardized value of None. This ensures that every record conveys meaningful feedback, rather than passing empty values to the model.
2. Preserving Linguistic Uniformity: Non-English text is filtered using the lingua-language-detector library to ensure linguistic consistency, mitigate sentiment analysis distortions, and align with the study's English-language focus.
3. Unicode Normalization and Encoding Conversion: Artifacts like special characters/symbols obscure textual clarity. To ensure analytical consistency, these were normalized to standardized Unicode using tool like the Unicode Converter[5] or systematically removed from the corpus.
4. Emoji-to-Text: Recognizing emojis' growing role in conveying sentiment and context within online feedback, we substituted each with textual descriptions, preserving emotional tone and contextual meaning rather than removing them.
5. Content Moderation for Ethical AI Practices: With the aid of *GPT-4o*, harmful or inappropriate content was identified and removed, maintaining dataset integrity while aligning with ethical AI standards and LLM content policies.

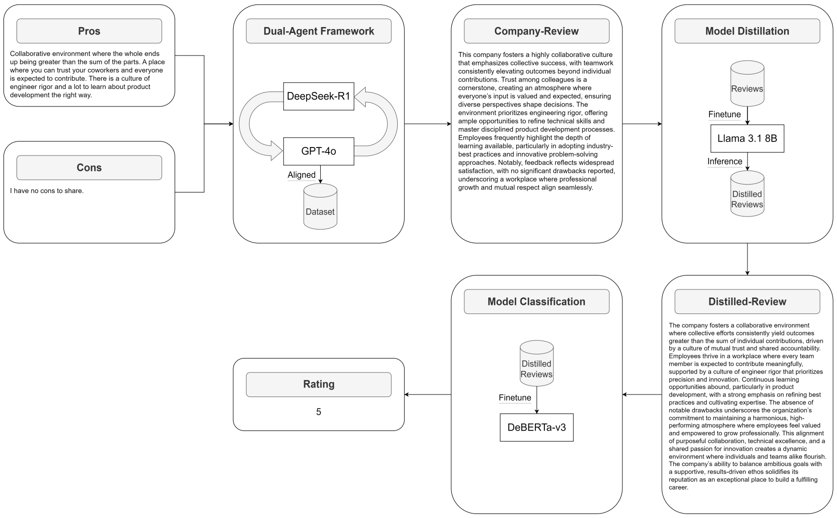
Following rigorous refinements, our dataset emerged as a robust collection of 20K records—blending 5% reviews from AmbitionBox with 95% reviews from Glassdoor. It was then methodically divided, allocating (70%) 14K records for training and splitting the remaining (30%) 6K evenly between validation and testing ensuring a balanced foundation for robust model development.

[2] `https://www.kaggle.com/datasets/saqlainrehan/employeesreviews-dataset`

[3] `https://www.glassdoor.com/Reviews/index.htm`

[4] `https://www.ambitionbox.com/list-of-companies?campaign=desktop_nav`

[5] `https://unicode.scarfboy.com`



Collaborative environment where the whole ends up being greater than the sum of the parts. A place where you can trust your coworkers and everyone is expected to contribute. There is a culture of engineer rigor and a lot to learn about product development the right way.

I have no cons to share.

This company fosters a highly collaborative culture that emphasizes collective success, with teamwork consistently elevating outcomes beyond individual contributions. Trust among colleagues is a cornerstone, creating an atmosphere where everyone's input is valued and expected, ensuring diverse perspectives shape decisions. The environment prioritizes engineering rigor, offering ample opportunities to refine technical skills and master disciplined product development processes. Employees frequently highlight the depth of learning available, particularly in adopting industry-best practices and innovative problem-solving approaches. Notably, feedback reflects widespread satisfaction, with no significant drawbacks reported, underscoring a workplace where professional growth and mutual respect align seamlessly.

The company fosters a collaborative environment where collective efforts consistently yield outcomes greater than the sum of individual contributions, driven by a culture of mutual trust and shared accountability. Employees thrive in a workplace where every team member is expected to contribute meaningfully, supported by a culture of engineer rigor that prioritizes precision and innovation. Continuous learning opportunities abound, particularly in product development, with a strong emphasis on refining best practices and cultivating expertise. The absence of notable drawbacks underscores the organization's commitment to maintaining a harmonious, high-performing atmosphere where employees feel valued and empowered to grow professionally. This alignment of purposeful collaboration, technical excellence, and a shared passion for innovation creates a dynamic environment where individuals and teams alike flourish. The company's ability to balance ambitious goals with a supportive, results-driven ethos solidifies its reputation as an exceptional place to build a fulfilling career.

Fig. 2: The architecture depicts the end-to-end pipeline, integrating a dual-agent framework for review generation, model distillation for efficiency, and classification for rating prediction

4 Experiment Results

4.1 Laying the Groundwork

Before embarking on the fine-tuning of LLMs, we laid a solid groundwork by leveraging cutting-edge ML techniques to benchmark accuracy and scalability. Our journey began with a hybrid approach, blending TFIDF with K-Nearest Neighbors (KNN), RF, NB, LR, and MLP, where the TFIDF- MLP pair emerged as the top performer attaining an accuracy rate of 46.08% and an F1 score of 45.30%.. This was followed by a second hybrid approach, pairing 6B-token pre-trained Global Vectors for Word Representation (GloVe) with KNN, RF, NB, and LR, with the GloVe-LR pair proving most effective achieving an accuracy rate of 42.35% and F1 score of 41.67%. We then shifted our focus to BiLSTM, conducting two experiments shown in Fig.3: the first, inspired by a similar study [1], aimed to implement its method but faced overfitting (achieving 37% F1 and Accuracy) in our dataset, which could be attributed to the fact that the approach in the aforementioned study was designed for a different dataset. Undeterred, we fine-tuned BiLSTM on our dataset in the second experiment, achieving convergence and outperforming all previous methods. Further validation came through Zero-Shot and Few-Shot testing with GPT-4o, where Few-Shot not only surpassed Zero-Shot but also edged out BiLSTM by a narrow margin in terms of accuracy and F1 score, both of which reached 50%. With these benchmarks firmly in place, we swiftly transition to optimizing LLMs for improved performance and efficiency using model distillation.

4.2 Model Distillation Evaluation

In our dual-agent framework, we integrated different variants of DeepSeek (R1 and V3) as composers, with GPT-4o acting as the critic, and assessed performance at temperatures of 1 and 0.5. Notably, DeepSeek R1 and GPT-4o operating at 0.5 temperature

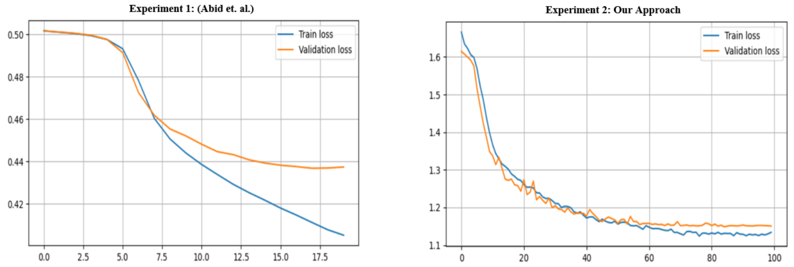


Fig. 3: Benchmarking BiLSTM: Experiment 1 used Abid et al.'s (2023) parameters, while Experiment 2 (our refined approach) achieved 50% accuracy and 49% F1

yielded the most balanced performance, effectively harmonizing creative expression with adherence to guidelines, and generating refined reviews suitable for subsequent model distillation. Building upon these findings, we extended our experiments to model distillation by evaluating multiple LLMs of varying scales—LLaMA 3.2 1B, LLaMA 3.1 8B, Qwen 2.5 1.5B, and Qwen 2.5 7B—under standardized conditions (six epochs, cosine learning rate scheduler, LoRA with rank 64, alpha 128, dropout 0.1, and gradient check-pointing). Model-specific hyper-parameters, such as learning rate and warmup steps, were fine-tuned for optimal performance. Within the LLaMA family, LLaMA 3.1 8B outperformed LLaMA 3.2 1B—achieving a validation loss of 1.1079 using a learning rate of 1e-4 and 500 warmup steps—while Qwen 2.5 7B attained a validation loss of 1.2091 using a learning rate of 8e-5 and 500 warmup steps. Both LLaMA 3.1 8B and Qwen 2.5 7B were further evaluated using 4-bit quantization. As depicted in Fig. 4, LLaMA 3.1 8B, emerged as the most effective model with the lowest loss, likely due to its larger capacity, advanced architecture, and greater suitability for the task. Notably, all models exhibited stable convergence, with no significant performance gains beyond the sixth epoch. Refer to Table 1 for a comparative performance summary.

Subsequently, we employed two variations of BERTScore, distilbert-base-uncased and deberta-xlarge-mnli to assess model performance. Interestingly, smaller models such as LLaMA 3.2 1B marginally outperformed their larger counterparts, likely due to the latter's tendency to drift off-topic toward the end of responses. However, after a universal post-processing, LLaMA 3.1 8B emerged as the leader across all tested models, achieving the highest BERTScores of 0.8641 and 0.7438 with distilbert-base-uncased and deberta-xlarge-mnli respectively, with its quantized version closely trailing behind, demonstrating efficient, scalable performance optimized for resource-efficient deployment. A similar trend was observed in Qwen 2.5 7B, the quantized Qwen 2.5 7B marginally outperformed its non-quantized version prior post-processing. Quantization's dual role as a stabilizer and regularizer [20, 13]—applied directly to the fine-tuned model—resolved Qwen's initial training instability, enhancing output consistency and BERTScore alignment. This step laid the foundation for classification using distilled LLaMA-based model outputs.

4.3 Model Classification Evaluation

We evaluated classification tasks on reviews generated by the distilled LLaMA 3.1 8B and its quantized version, both of which achieved the highest BERTScore, demonstrat-

Table 1: Evaluation of fine-tuning different open source LLMs

Model	Training Loss (CE)	Val. Loss (CE)	GPUs	Time
Llama 3.1 8B	1.0423	1.1079	4 × A100	∼2.5 h
Llama 3.2 1B	1.3178	1.3657	4 × A100	∼0.5833 h
Qwen 2.5 7B	1.1863	1.2091	4 × A100	∼2.2 h
Qwen 2.5 1.5B	1.2559	1.3199	4 × A100	∼0.75 h

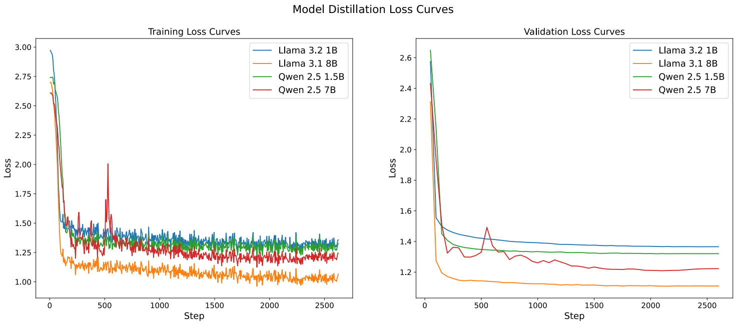


Fig. 4: A set of models was chosen for distillation; the graph above indicates that Lama 3.1 8B seems to have the least loss. Notably, Qwen 2.5 7B experienced a spike during training that eventually levelled off

ing superior alignment with the teacher model among all distilled models. To assess performance, we fine-tuned RoBERTa-Large and DeBERTa-v3 Large on these outputs. LLaMA 3.1 8B emerged as the top performer across both models, slightly edging out its quantized counterpart. Consistent hyper-parameters were maintained: 4 epochs, a cosine learning rate scheduler, and a batch size of 16, both RoBERTa large and DeBERTa-v3 achieved optimal performance on LLaMA 3.1 8 with a 3e-6 learning rate and 0.1 warmup ratio. The strongest combination was LLaMA 3.1 8B with DeBERTa-v3, achieving 55.43% accuracy, 56.32% Precision, 55.43% Recall and a 55.73% F1 score, see Table 2. Interestingly, RoBERTa thrived at the extremes, excelling in classes 1 and 5, yet struggled in the mid-range (2 and 4). In contrast, DeBERTa-v3 held its own at the extremes and slightly outperformed RoBERTa in the mid-range, showcasing its adaptability to nuanced classifications, refer to Fig. 5. This advantage is attributed to its superior attention mechanism and enhanced positional encoding, allowing it to capture complex linguistic patterns more effectively than RoBERTa. Nevertheless, additional tuning is necessary to improve both accuracy and F1 score.

4.4 Fair Attribution in AI: Shapley-Based Analysis of DeBERTa Predictions

Explainable AI plays a pivotal role in bridging the gap between complex model decisions and human understanding, and Shapley values stand at the forefront of this effort by ensuring fair attribution of each feature's contribution to predictions. In our study, we integrated the output of our best-performing model, DeBERTa-v3, into Shapley analysis and employed the one-vs-rest technique to transform multiclass classification into multiple binary settings. The waterfall plot proved instrumental in visualizing these attributions, suggesting that human evaluations may exhibit bias or inconsistency. For

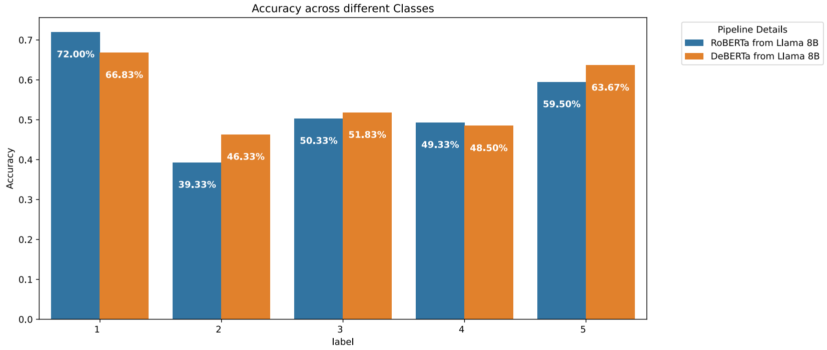


Fig. 5: The graph of accuracy metrics reveals that LLaMA-DeBERTa achieves stronger overall performance, whereas LLaMA-RoBERTa struggles specifically with mid-range labels (2–4)

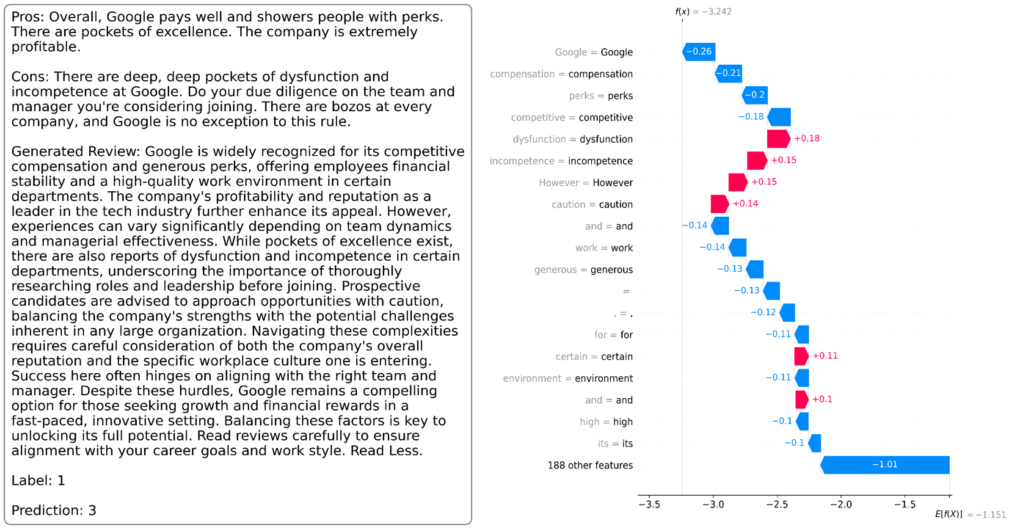
Table 2: Comparative classifier performance metrics, with a focus on the delta between baseline and experimental models.

Classifier	LLM	Accuracy	Precision	Recall	F1
Baseline (BiLSTM)	-	0.500	0.490	0.500	0.490
RoBERTa	Llama 8B	0.541 (+8.20%)	0.550 (+12.29%)	0.541 (+8.20%)	0.543 (+10.73%)
	Quantized Llama 8B	0.541 (+8.20%)	0.541 (+10.31%)	0.541 (+8.20%)	0.539 (+9.95%)
DeBERTa	Llama 8B	0.554 (+10.87%)	0.563 (+14.94%)	0.554 (+10.87%)	0.557 (+13.74%)
	Quantized Llama 8B	0.548 (+9.53%)	0.570 (+16.31%)	0.548 (+9.53%)	0.553 (+12.81%)

instance, as shown in Fig 6, words like compensation, perks, competitive, and generous—which carry an inherently optimistic and positive connotation—systematically led the model to assign a higher organizational score of 3—most likely in line with objective human grading—despite the initial rating appeared unjustifiably low at 1. In contrast, another case demonstrated an employee who provided no complaints yet rated their company a 4. The model, however, predicted a 5, which aligned more logically with the positive sentiment expressed, see Fig. 7. These findings underscore the value of AI-driven sentiment analysis in mitigating subjective inconsistencies and promoting fairness in organizational assessments.

5 Conclusion and Future Work

This study charts a transformative path from foundational ML techniques to the innovative application of LLMs, culminating in a robust framework for translating organizational sentiment into quantifiable insight. Beginning with an exploration of state-of-the-art ML methods, we established the groundwork for capturing the nuanced spectrum of employee perceptions. Building on this foundation, we introduced a dual-agent framework architecture, to synthesize authentic, context-rich company reviews. These outputs were subsequently distilled into an efficient, leaner model, enabling scalable



Pros: Overall, Google pays well and showers people with perks. There are pockets of excellence. The company is extremely profitable.

Cons: There are deep, deep pockets of dysfunction and incompetence at Google. Do your due diligence on the team and manager you're considering joining. There are bozos at every company, and Google is no exception to this rule.

Generated Review: Google is widely recognized for its competitive compensation and generous perks, offering employees financial stability and a high-quality work environment in certain departments. The company's profitability and reputation as a leader in the tech industry further enhance its appeal. However, experiences can vary significantly depending on team dynamics and managerial effectiveness. While pockets of excellence exist, there are also reports of dysfunction and incompetence in certain departments, underscoring the importance of thoroughly researching roles and leadership before joining. Prospective candidates are advised to approach opportunities with caution, balancing the company's strengths with the potential challenges inherent in any large organization. Navigating these complexities requires careful consideration of both the company's overall reputation and the specific workplace culture one is entering. Success here often hinges on aligning with the right team and manager. Despite these hurdles, Google remains a compelling option for those seeking growth and financial rewards in a fast-paced, innovative setting. Balancing these factors is key to unlocking its full potential. Read reviews carefully to ensure alignment with your career goals and work style. Read Less.

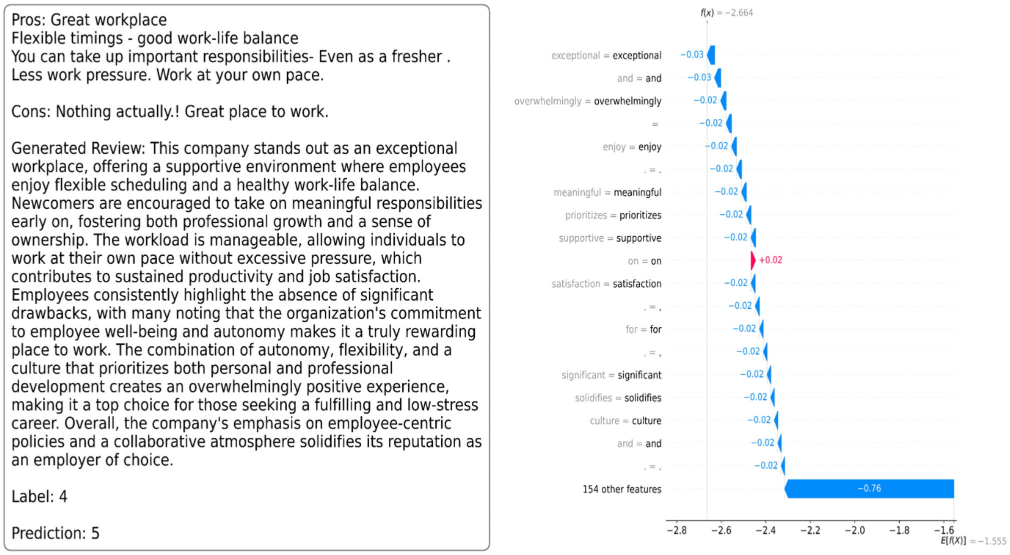
Label: 1

Prediction: 3

Fig. 6: The SHAP waterfall plot highlights word contributions, with red indicating support for class 1 and blue representing influence toward classes other than class 1. The prominent blue contributions suggest that predicated 3 may better reflect the employee feedback than the label of class 1

deployment without sacrificing depth. To anchor abstract sentiment into actionable insights, the distilled outputs were classified onto a 5-point Likert scale. Recognizing the risks of subjective bias, we embedded Explainable AI (XAI) principles, leveraging SHAP values to surface hidden patterns in the classification process. Through this, we were able to identify bias in employee ratings and detect subjective inconsistencies between the textual sentiment and numerical scores. These insights will not only strengthen model interpretability but will also contribute to more equitable and transparent organizational assessments. Beyond corporate and organizational research, this framework has the potential to propel interdisciplinary research: In education technology, it facilitates automated, granular analysis of student feedback to enhance learning experiences, and also informs Marketing and Consumer Behavior strategies by capturing multidimensional audience sentiment to optimize brand resonance.

Future work will focus on addressing key challenges in model distillation, including persistent hallucinations and inherited flaws from the teacher model. Enhanced tokenization methods from recent Hugging Face updates will be used to improve end-of-sequence (EOS) prediction and reduce irrelevant text, while reinforcement learning (RL) will be applied to penalize misaligned outputs and strengthen contextual coherence. Ethical safeguards will ensure transparency in rating methodologies through rigorous bias audits to mitigate the risks posed by hallucinations and biases during company ratings, while also engaging stakeholders to align model behavior with equitable standards. Beyond these steps, hyper-parameter tuning will be explored to optimize performance across diverse organizational settings. Together, these efforts aim to improve robustness, reduce error propagation, and better align the student model with human evaluative standards grounded in fairness, accountability, and transparency, thereby improving both performance and reliability in downstream applications.



Pros: Great workplace
Flexible timings - good work-life balance
You can take up important responsibilities- Even as a fresher .
Less work pressure. Work at your own pace.

Cons: Nothing actually.! Great place to work.

Generated Review: This company stands out as an exceptional workplace, offering a supportive environment where employees enjoy flexible scheduling and a healthy work-life balance. Newcomers are encouraged to take on meaningful responsibilities early on, fostering both professional growth and a sense of ownership. The workload is manageable, allowing individuals to work at their own pace without excessive pressure, which contributes to sustained productivity and job satisfaction. Employees consistently highlight the absence of significant drawbacks, with many noting that the organization's commitment to employee well-being and autonomy makes it a truly rewarding place to work. The combination of autonomy, flexibility, and a culture that prioritizes both personal and professional development creates an overwhelmingly positive experience, making it a top choice for those seeking a fulfilling and low-stress career. Overall, the company's emphasis on employee-centric policies and a collaborative atmosphere solidifies its reputation as an employer of choice.

Label: 4

Prediction: 5

Fig. 7: The Shapley waterfall plot visualizes word contributions, where red represents class 4 and blue denotes influence toward classes other than class 4. The prevalence of blue indicates that the predicated class 5 is a more suitable classification than the label of class 4

References

[1] Muhammed Yaseen Morshed Adib et al. "BiLSTM-ANN Based Employee Job Satisfaction Analysis from Glassdoor Data Using Web Scraping". In: *Procedia Computer Science* 222 (2023), pp. 1–10. DOI: `https://doi.org/10.1016/j.procs.2023.08.139`.

[2] Rajiv Bajpai et al. "Aspect-Sentiment Embeddings for Company Profiling and Employee Opinion Mining". In: *Computational Linguistics and Intelligent Text Processing*. Ed. by Alexander Gelbukh. Cham: Springer Nature Switzerland, 2023, pp. 142–160.

[3] Vimala Balakrishnan et al. "A deep learning approach in predicting products' sentiment ratings: a comparative analysis". In: *The Journal of Supercomputing* 78.5 (2022), pp. 7206–7226.

[4] DeepSeek-AI et al. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. 2025. arXiv: `2501.12948 [cs.CL]`. URL: `https://arxiv.org/abs/2501.12948`.

[5] Aya E Fouda et al. "Sentiment Analysis On Arabic Companies Reviews". In: *2024 6th International Conference on Computing and Informatics (ICCI)*. IEEE. 2024, pp. 418–428.

[6] Ankur Joshi et al. "Likert scale: Explored and explained". In: *British Journal of Applied Science & Technology* 7.4 (2015), p. 397. DOI: `https://doi.org/10.9734/BJAST/2015/14975`.

[7] Waleed Kareem and Noorhan Abbas. "Fighting lies with intelligence: Using large language models and chain of thoughts technique to combat fake news". In:

International Conference on Innovative Techniques and Applications of Artificial Intelligence. Springer Nature Switzerland, 2023, pp. 254–257. DOI: https://doi.org/10.1007/978-3-031-47994-6_24.

[8] Shih-Wen Ke, Chih-Fong Tsai, and Yi-Jun Chen. "Managing Emotion In The Workplace: An Empirical Study With Enterprise Instant Messaging". In: *Applied Artificial Intelligence* 38.1 (2024), p. 2. DOI: https://doi.org/10.1080/08839514.2023.2297518.

[9] Stanislava Kozakijevic et al. "Machine Learning for Company Review Sentiment Analysis Interpretation". In: *International Conference on Multi-Strategy Learning Environment*. Springer. 2024, pp. 647–659. DOI: https://doi.org/10.1007/978-981-97-1488-9_47.

[10] Dawei Li et al. *Preference Leakage: A Contamination Problem in LLM-as-a-judge*. 2025. arXiv: 2502.01534 [cs.LG]. URL: https://arxiv.org/abs/2502.01534.

[11] Chapman J Lindgren et al. "Sentiment Analysis for Organizational Research". In: *Stress and Well-being at the Strategic Level*. Emerald Publishing Limited, 2023, pp. 95–117.

[12] Aman Madaan et al. "Self-Refine: Iterative Refinement with Self-Feedback". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 46534–46594.

[13] Markus Nagel et al. *A White Paper on Neural Network Quantization*. 2021. arXiv: 2106.08295 [cs.LG]. URL: https://arxiv.org/abs/2106.08295.

[14] Libo Qin et al. *Large Language Models Meet NLP: A Survey*. 2024. arXiv: 2405.12819 [cs.CL]. URL: https://arxiv.org/abs/2405.12819.

[15] Muhammad Saqlain Rehan et al. "Employees reviews classification and evaluation (ERCE) model using supervised machine learning approaches". In: *Journal of Ambient Intelligence and Humanized Computing* 13.6 (2022), pp. 3119–3136.

[16] Lois Rink, Job Meijdam, and David Graus. "Aspect-Based Sentiment Analysis for Open-Ended HR Survey Responses". In: *arXiv preprint arXiv:2402.04812* (2024), p. 1.

[17] Konstantinos I Roumeliotis, Nikolaos D Tselikas, and Dimitrios K Nasiopoulos. "LLMs in e-commerce: a comparative analysis of GPT and LLaMA models in product review evaluation". In: *Natural Language Processing Journal* 6 (2024), p. 100056.

[18] Ankit Taparia and Tanmay Bagla. "Sentiment analysis: predicting product reviews' ratings using online customer reviews". In: *Available at SSRN 3655308* (2020).

[19] Chen Wang et al. "A Novel Multi-Class Product Rating Prediction Model based on Enhanced Textrank Text Encoding and Human Psychology Simulation". In: *2024 IEEE International Conference on Industrial Technology (ICIT)*. IEEE. 2024, pp. 1–8.

[20] Hanwei Wu and Markus Flierl. “Vector Quantization-Based Regularization for Autoencoders”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.04 (Apr. 2020), pp. 6380–6387. DOI: `10.1609/aaai.v34i04.6108`.

[21] Hrithika Yadav, Kartik Dwivedi, and G Abirami. “Sentiment Analysis of company reviews using Machine Learning”. In: *2022 3rd International Conference for Emerging Technology (INCET)*. IEEE. 2022, pp. 1–5.

[22] Yongxiong Zhang and Liangming Wang. “Design of employee comment sentiment analysis platform based on AE-SVM algorithm”. In: *Journal of Physics: Conference Series*. Vol. 1575. 1. IOP Publishing. 2020, p. 4. DOI: `https://doi.org/10.1088/1742-6596/1575/1/012019`.